

Technology emerging from the DEEP & DEEP-ER projects

Estela Suarez
Jülich Supercomputing Centre

03.06.2016

DEEP

- **Cluster-E**
- Software
- Programn
- Energy ef
- Application
 - Co-desi
 - Evaluat
 - Code m

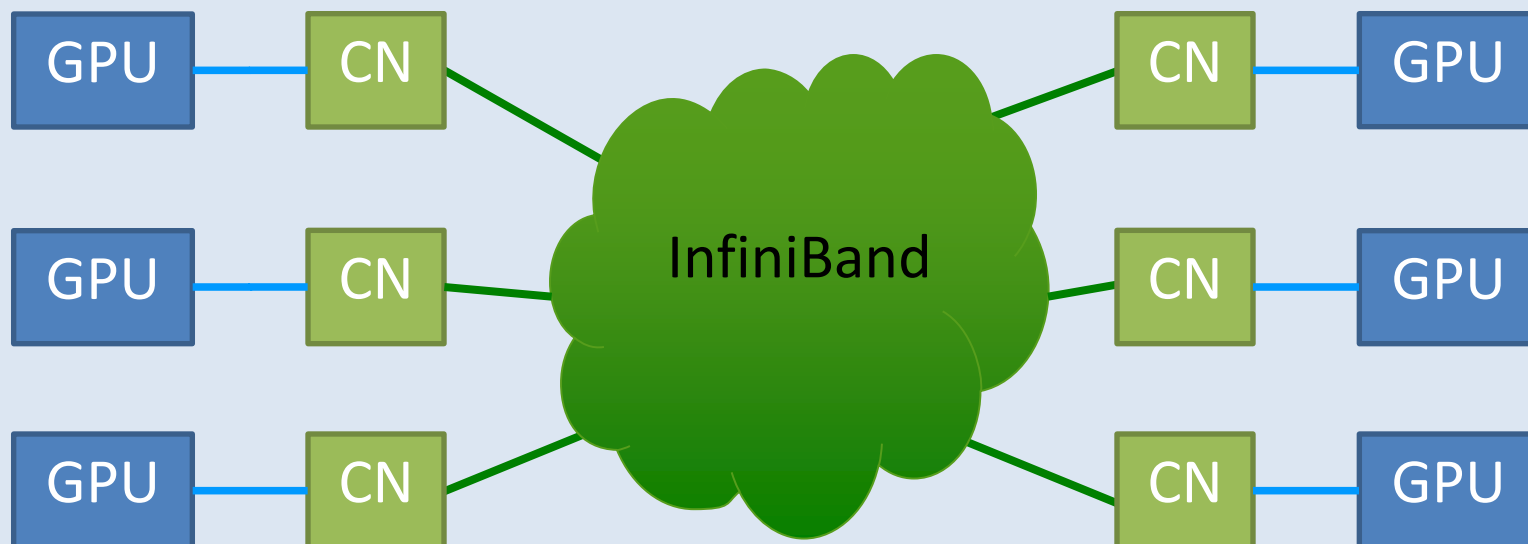
DEEP/-ER bring new technologies in:

Hardware
and
Software

hierarchy
ce **I/O**
ncy

stration
Code modernisation

CLUSTER-BOOSTER ARCHITECTURE

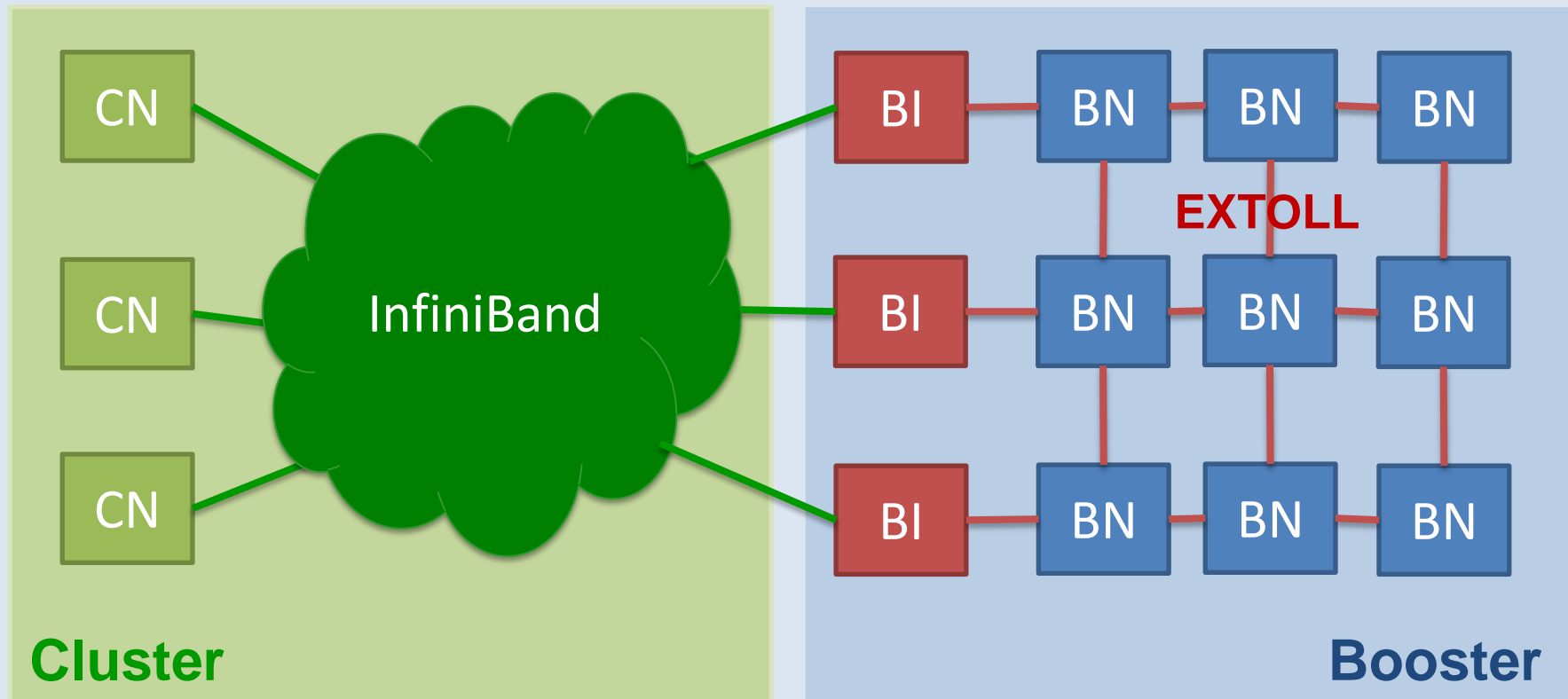


Flat topology

Simple management of
resources

Static assignment of
accelerators to CPUs

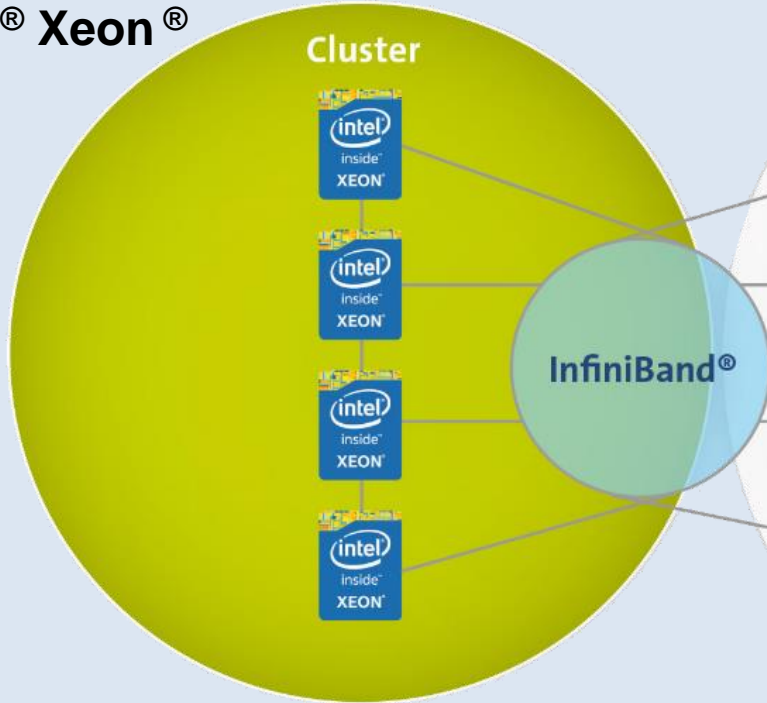
Accelerators cannot act
autonomously



Flexible assignment of resources (CPUs, accelerators)
Direct communication between accelerators
“Offload” of large and complex parts of applications

Intel® Xeon®

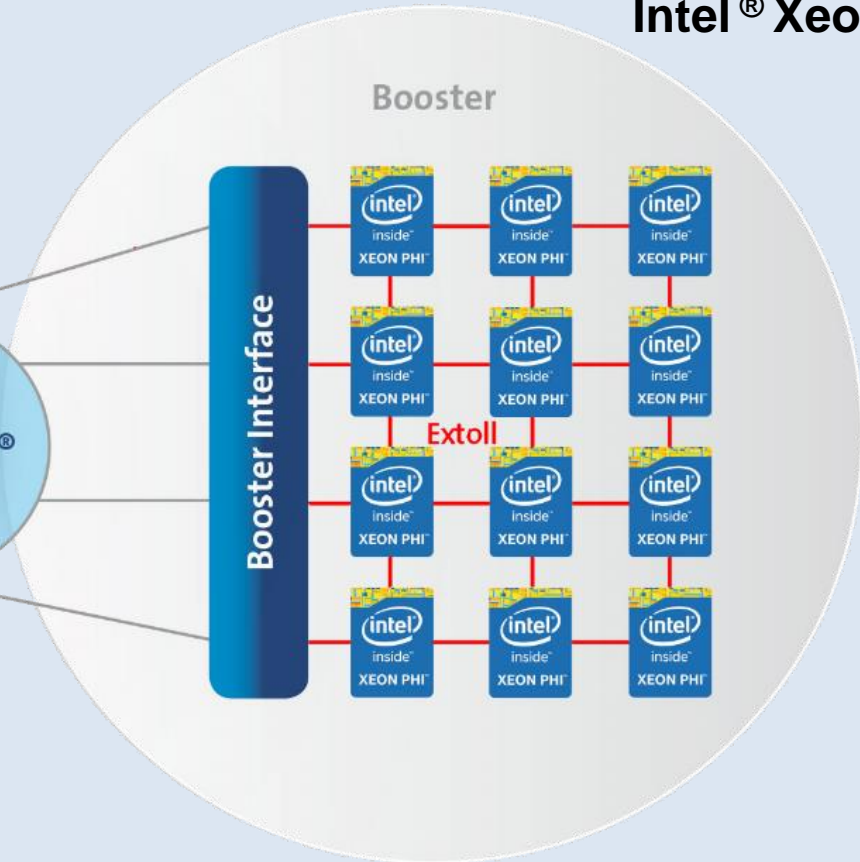
Cluster



LOW/MEDIUM
SCALABLE CODE

Intel® Xeon Phi™

Booster



HIGHLY
SCALABLE CODE

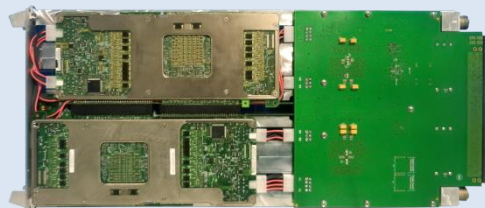
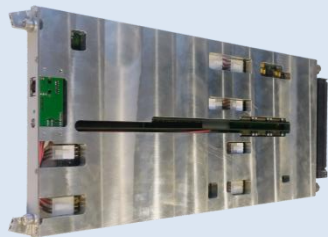
- Installed at JSC
- 1,5 racks
- 500 TFlop/s peak perf.
- 3.5 GFlop/s/W
- Water cooled

**Cluster
(128 Xeon)**

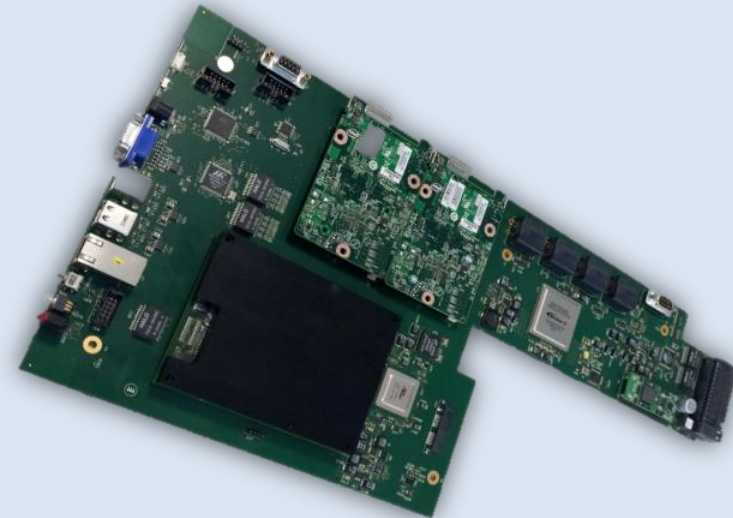


**Booster
(384 Xeon Phi
KNC)**

Node Card

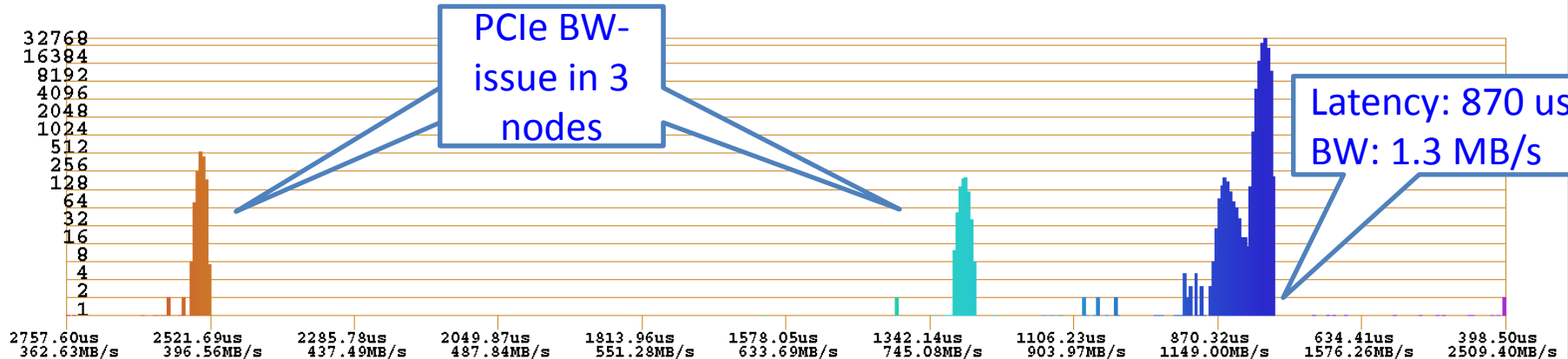


Interface Card



Booster measurements

MPI Linktest: ping-pong



length_of_message: 1048576 bytes (1024.00 KBytes)
 number_of_messages: 25
 Alltoall: 1
 Min Value: 398.5us (2509.40 MB/s)
 Max Value: 2757.6us (362.63 MB/s)
 Avg Value: 814.9us (1227.12 MB/s)

number_of_tasks: 384
 Execution order: Serial
 Mixing PE rank: No
 Alltoall Min Value: 616.1us (1 Byte)
 Alltoall Max Value: 5038.0us (1 Byte)
 Alltoall Avg Value: 795.1us (1 Byte)

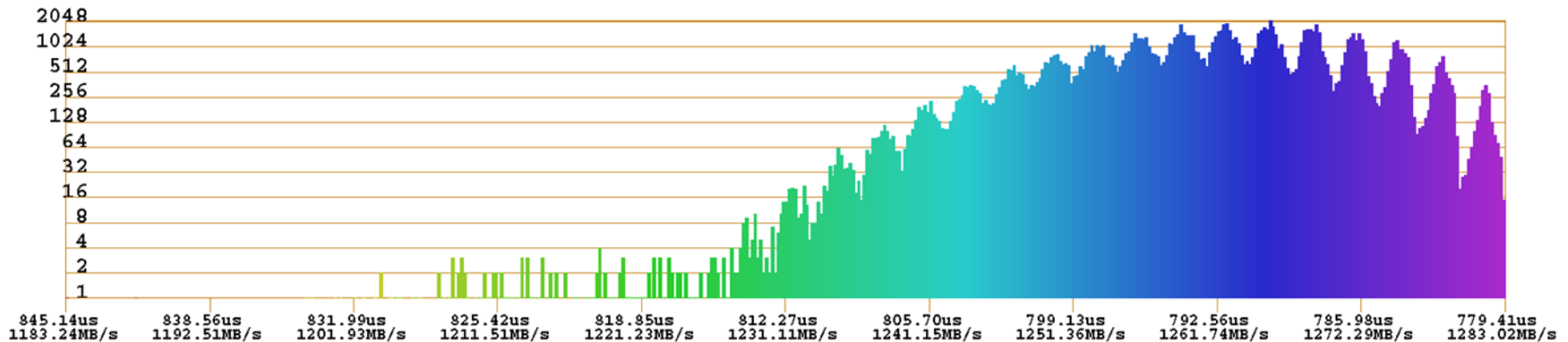
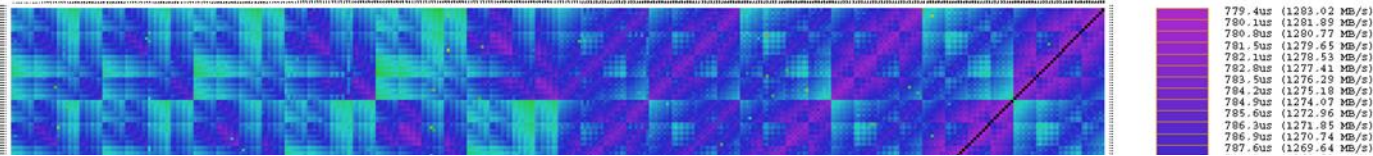
Report generated by FZJ Linktest Result Analyzer, Forschungszentrum Juelich GmbH



Booster Nodes (KNC) from 1 to 384

Booster measurements

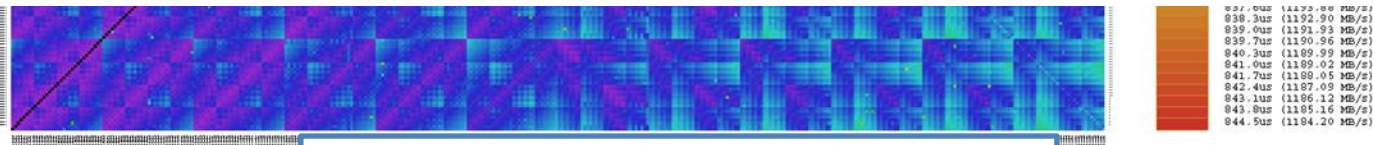
MPI Linktest: ping-pong



length_of_message:	1048576 bytes (1024.00 KBytes)	number_of_tasks:	384
number_of_messages:	40	Execution order:	Serial
Alltoall:	0	Mixing PE rank:	No
Min Value:	779.4us (1283.02 MB/s)		
Max Value:	845.1us (1183.24 MB/s)		
Avg Value:	792.7us (1261.47 MB/s)		

Stddev < 10%

Report generated by FZJ Linktest Result Analyzer, Forschungszentrum Juelich GmbH



Booster Nodes (KNC) from 1 to 384

Tourmalet PCI Express Board

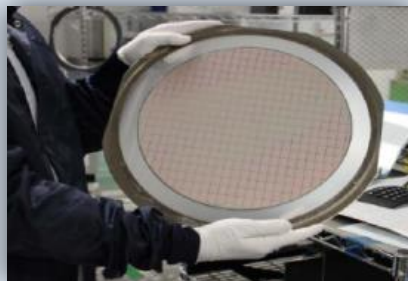


Main EXTOLL characteristics

- Direct network: no switches required
- Integrates network interface controller
- Supports 6+1 links
- Capable of tunneling PCIe (allows remote-booting KNC from the network)

Current (A3) version of EXTOLL ASIC

- 270 million transistors
- Link bandwidth: 100 G
- MPI latency: 850 ns
- MPI bandwidth: 8.5 GB/s
- Message rate: 70 million mgs/sec
- PCIe Gen3 x16



Tourmalet Chip and Wafer

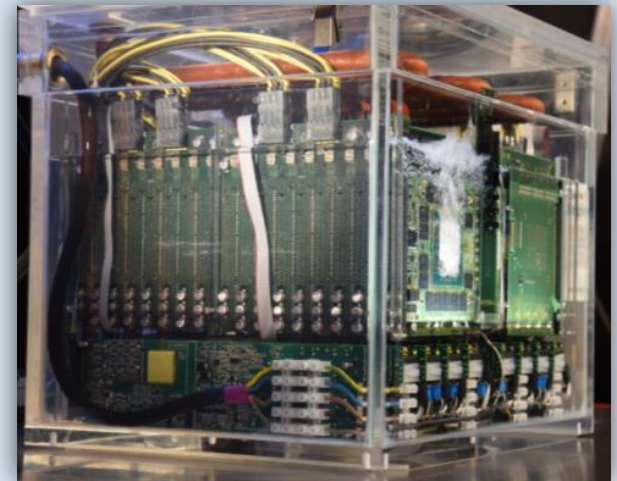
Alternative Booster implementation

- Interconnect EXTOLL **ASIC “Tourmalet”**
- 32 KNC-node system
- Implement $4 \times 4 \times 2$ topology, with Z dimension open



2-phase immersion cooling

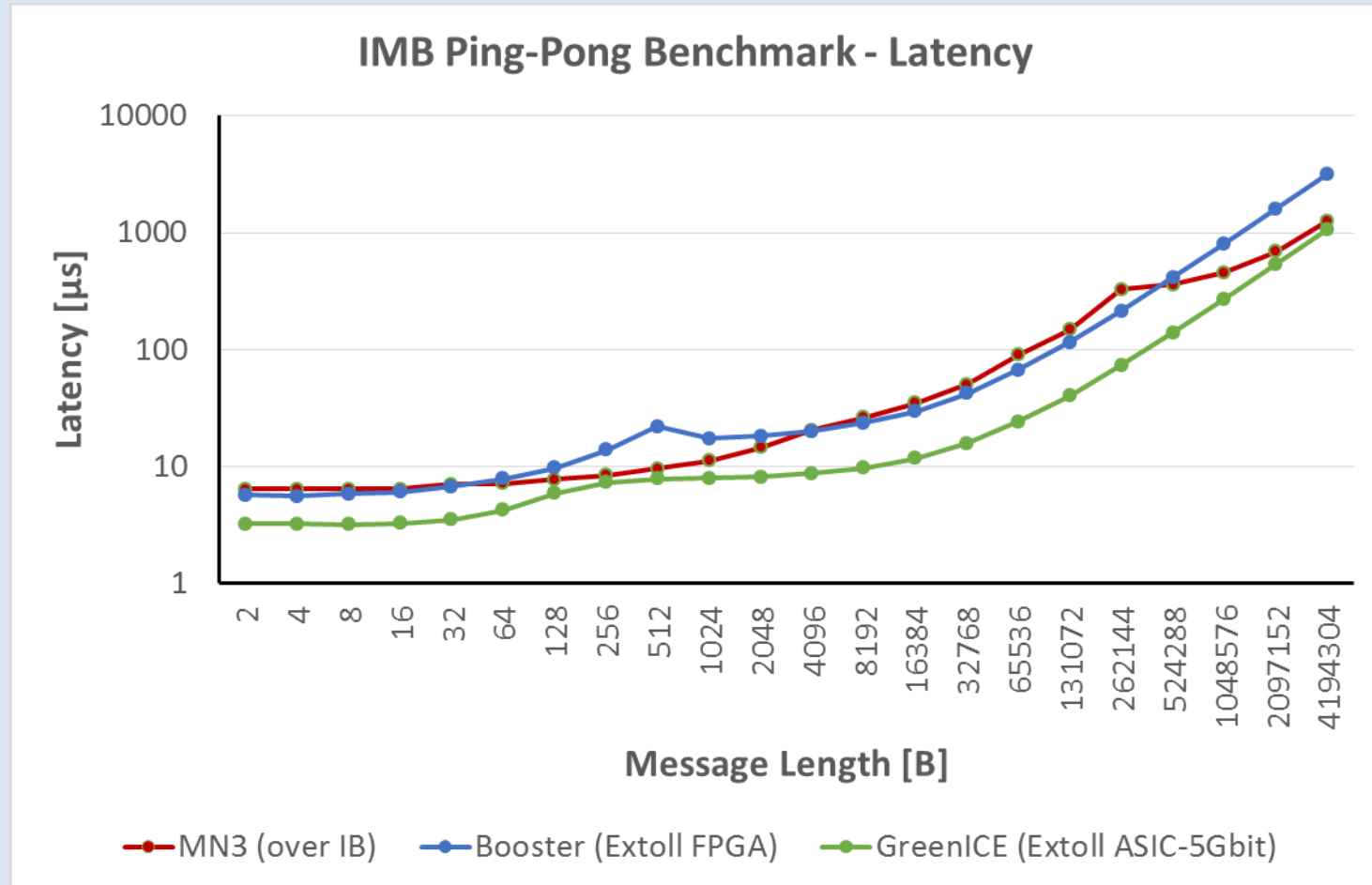
- NOVEC liquid from 3M
- Evaporates at about 50 degrees
- Condensates again in a water cooling pipe
- Allows very high-density integration



GreenICE Booster

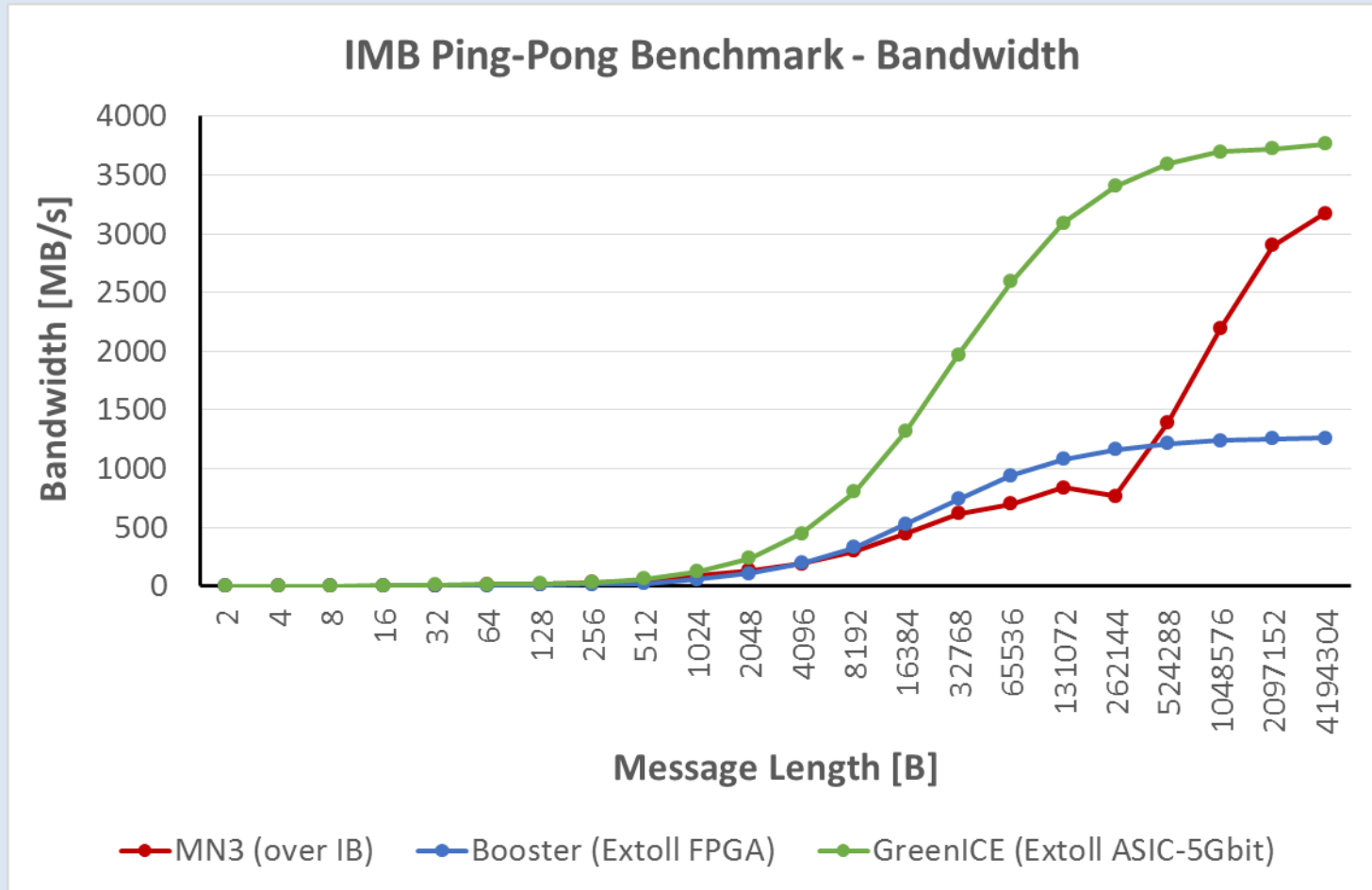
MPI performance

Latency



MPI performance

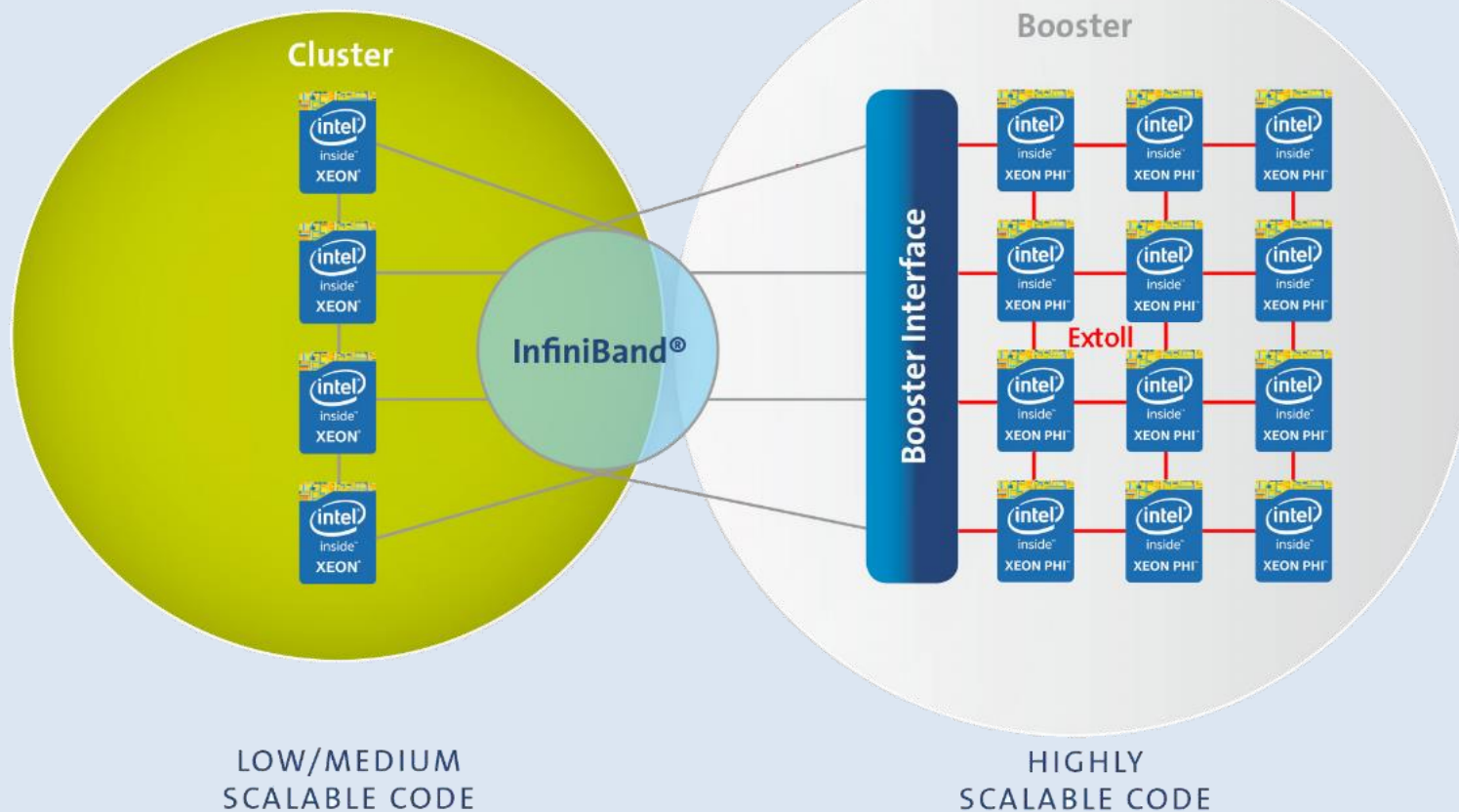
Bandwidth



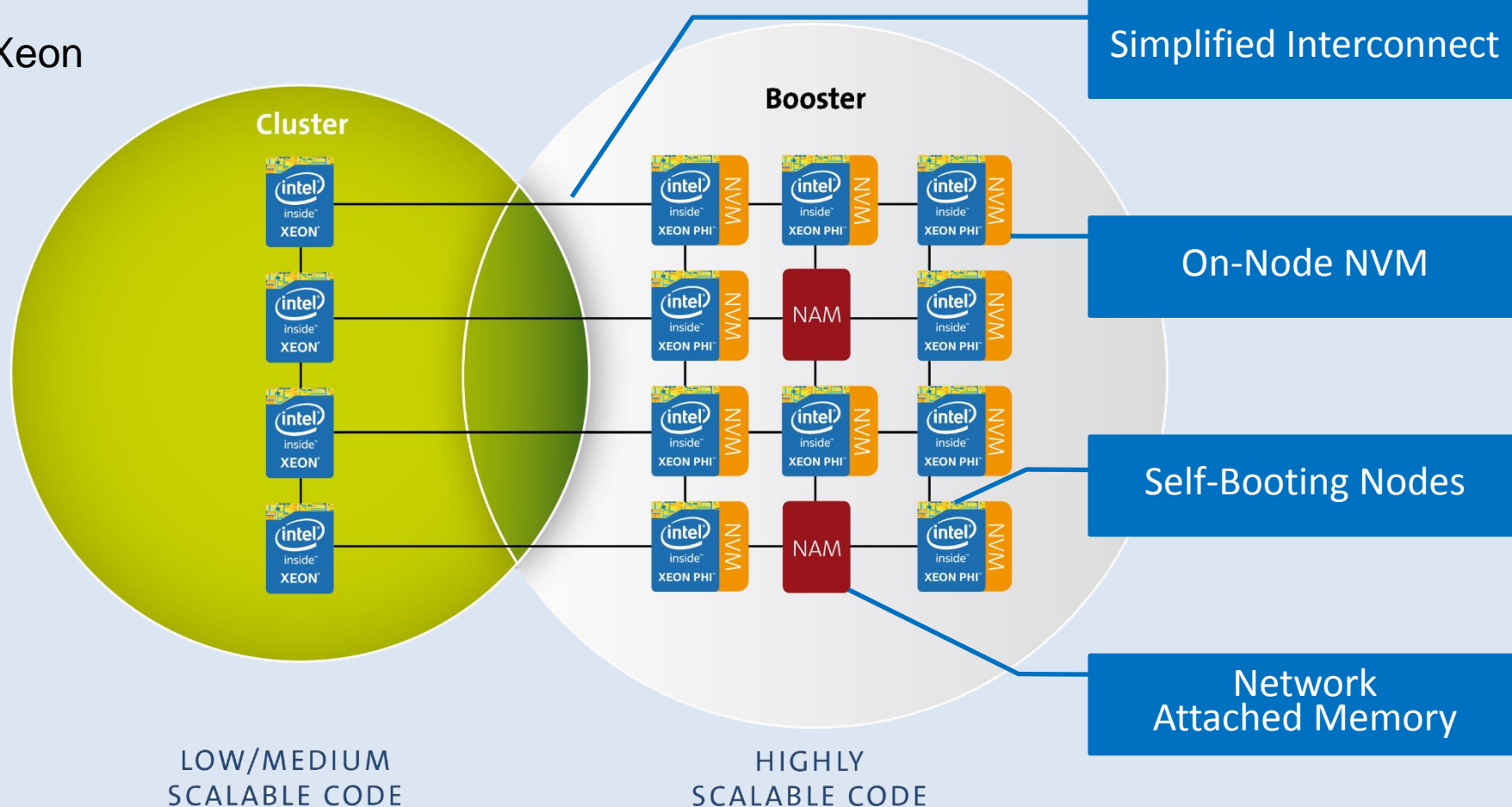
Factor of 3× achieved by EXTOLL TOURMALET (5Gbit/s version A2)

Xeon

Xeon Phi



Xeon



DEEP-ER Aurora Blade prototype



Eurotech's Aurora technology

Direct water cooled, high density

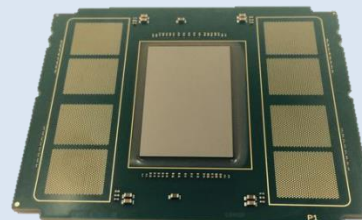
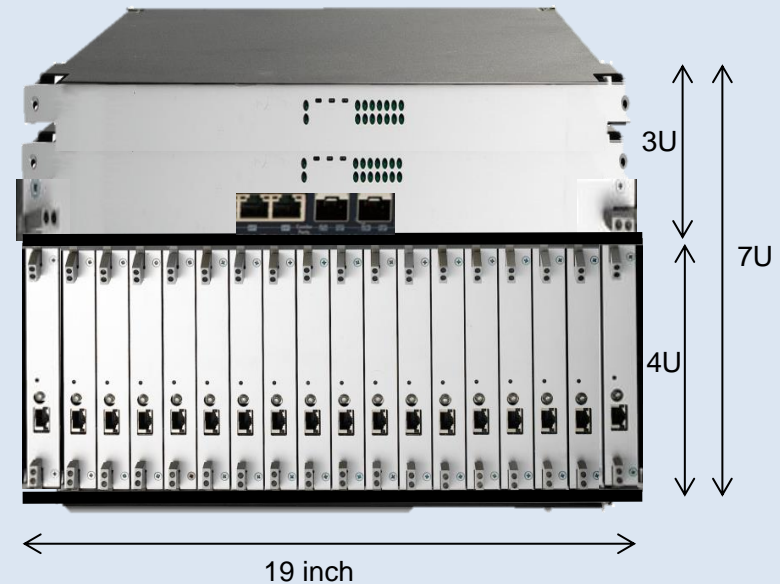


Aurora Blade DEEP-ER Booster
(in construction)

Rootcard
-18x EXTOLL
-18x NVMe

Chassis:
-18x KNL
-94GB Mem
-1x backplane

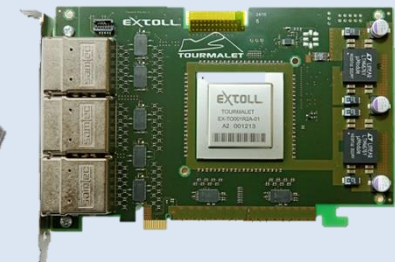
Aurora Blade Chassis



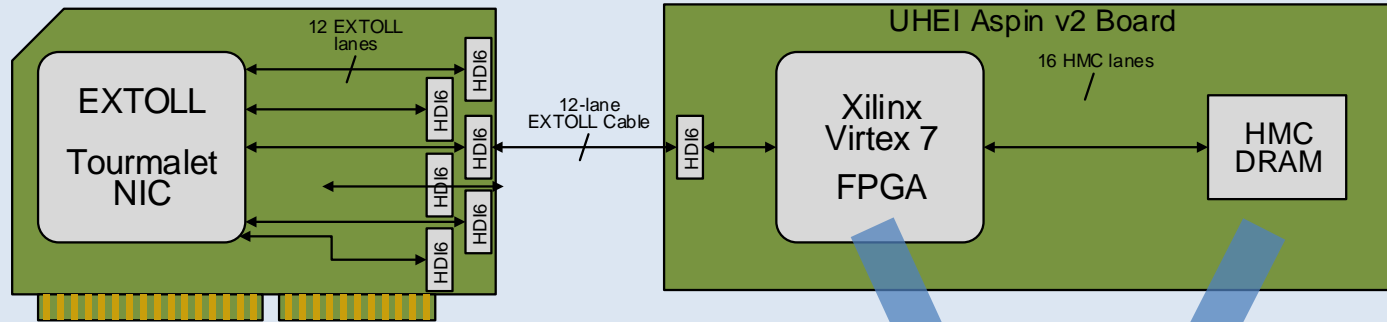
**Intel Xeon Phi
(KNL)**



NVMe



EXTOLL Tourmalet



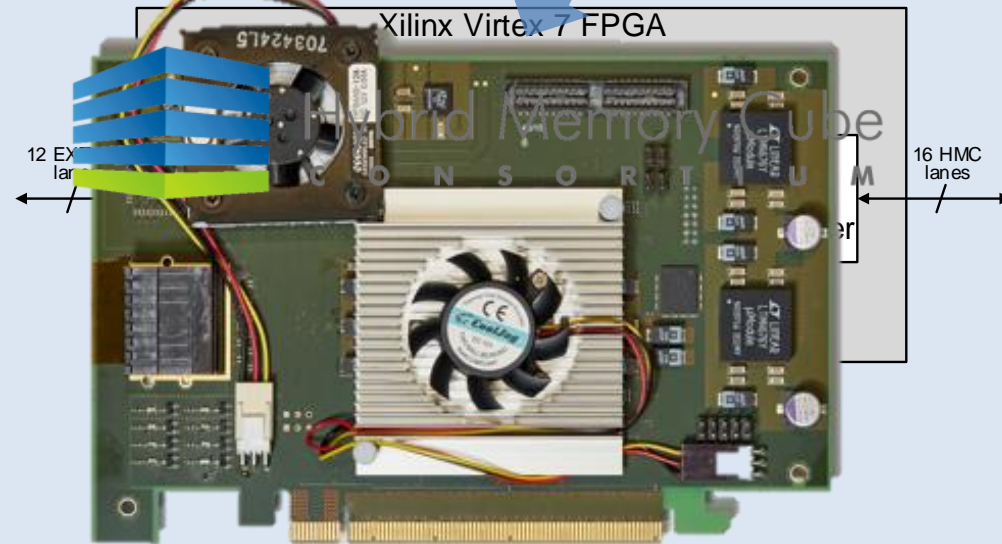
NAM architecture

- Xilinx Virtex 7 FPGA
- Hybrid Memory Cube (HMC)
 - Bandwidth HMC↔FPGA: 40+ GByte/s
 - HMC Controller: Open source development
- Attached to TOURMALET NIC

libNAM (libc based) for ease of use

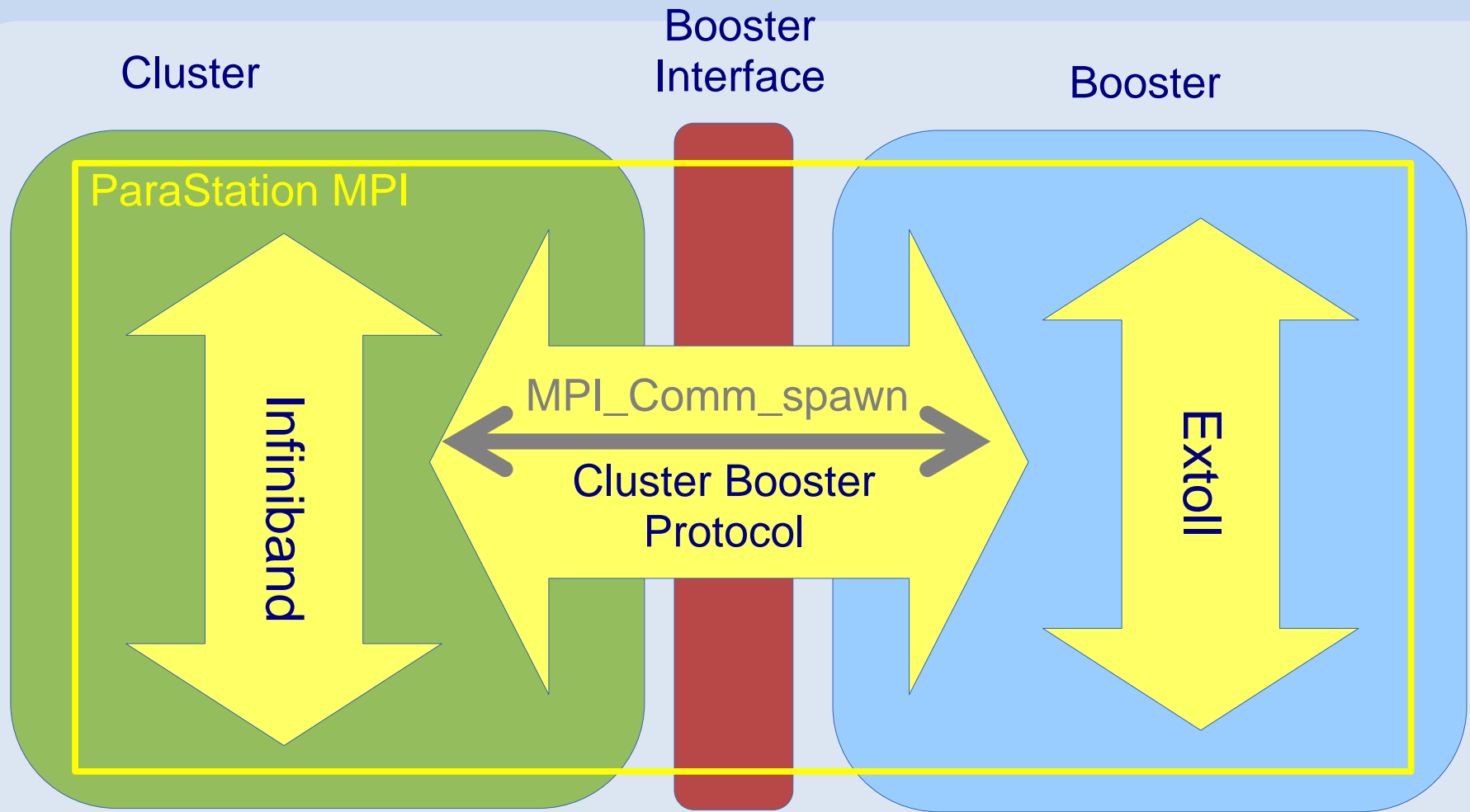
Use cases:

- global (shared) storage
- compute node for an X-OR C/R app
- “active memory”, etc.

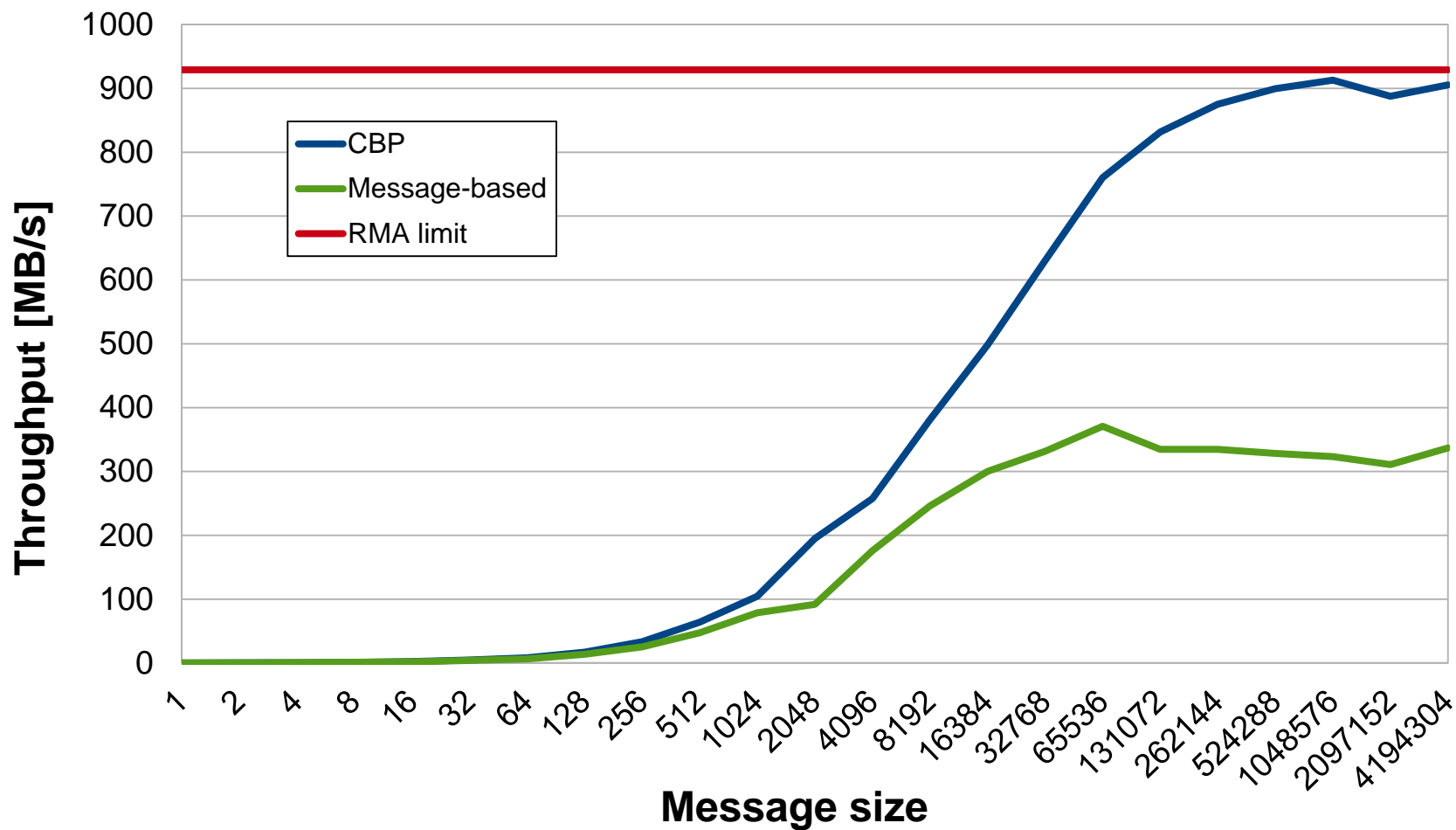


NAM Board

SOFTWARE



OmpSs on top of MPI provides pragmas to ease the offload process



Source code

```
int main(int argc, char *argv[]){
    /*...*/
    for(int i=0; i<3; i++){
        #pragma omp task in(...) out (...) onto (com, size*rank+1)
        foo_mpi(i, ...);}
}
```

Compiler

OmpSs Compiler

Application binaries

Cluster Executable

Booster Executable

DEEP Runtime

Cluster MPI

ParaStation Global MPI

DEEP Runtime

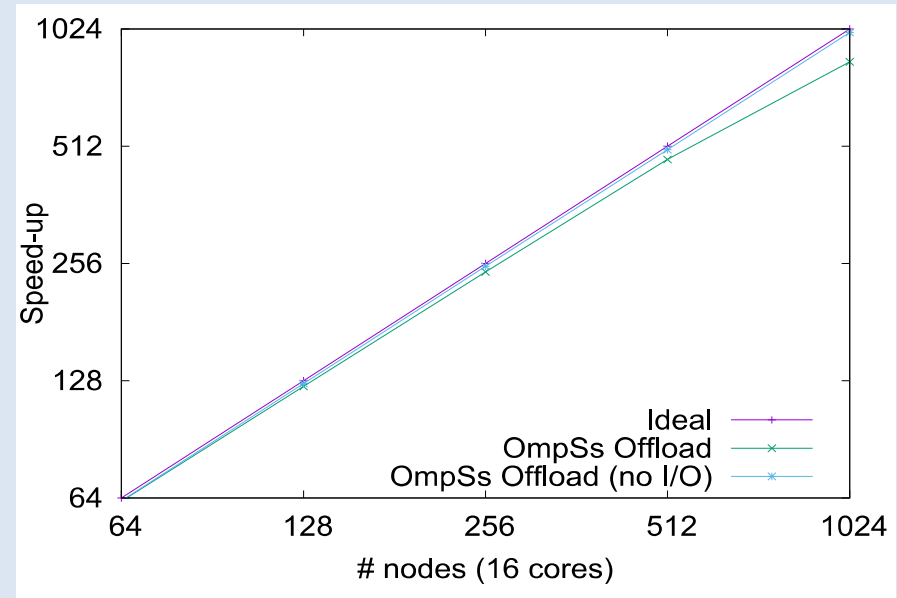
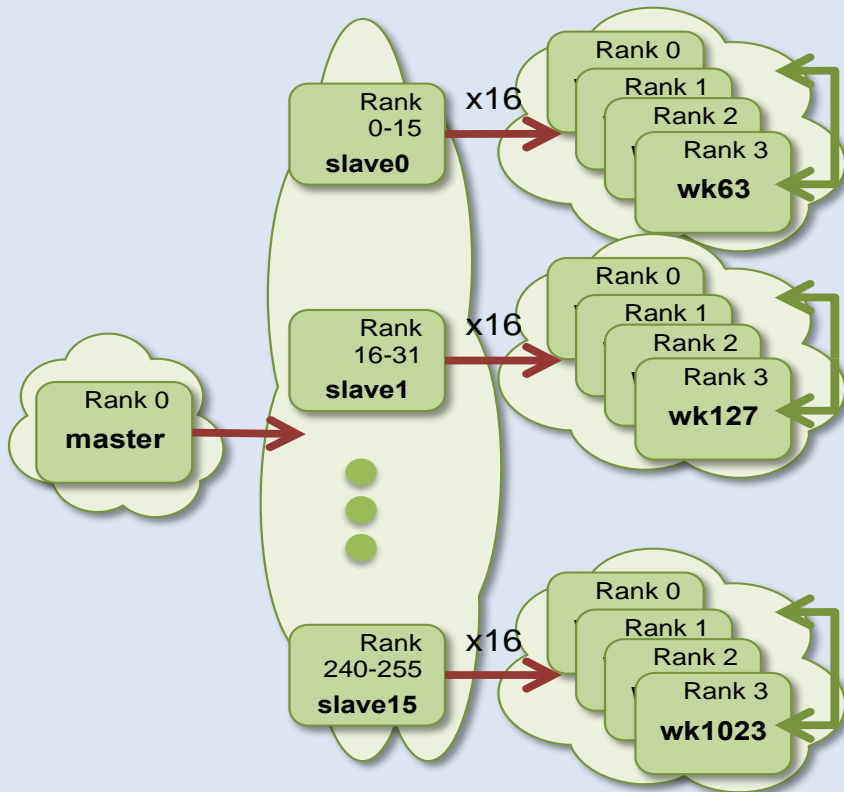
Booster MPI

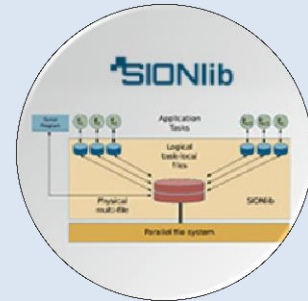
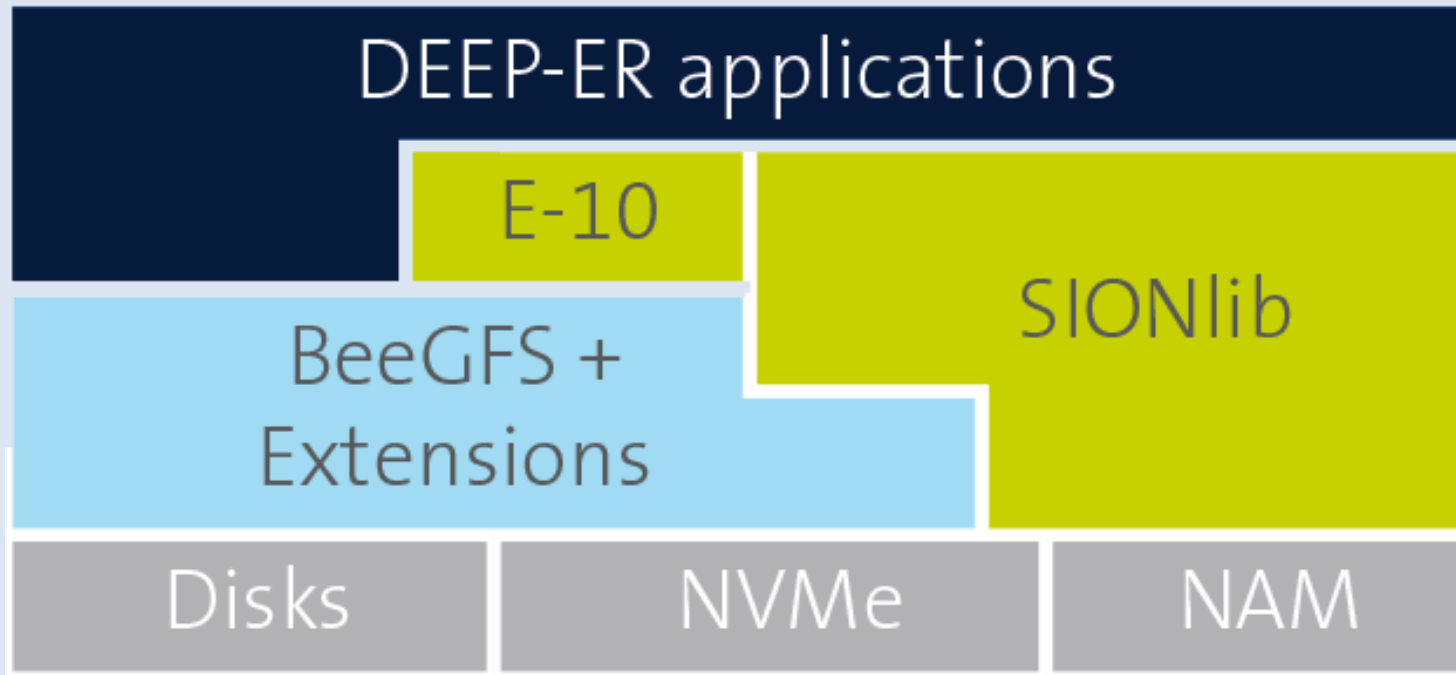
OmpSs Runtime

CLUSTER

BOOSTER

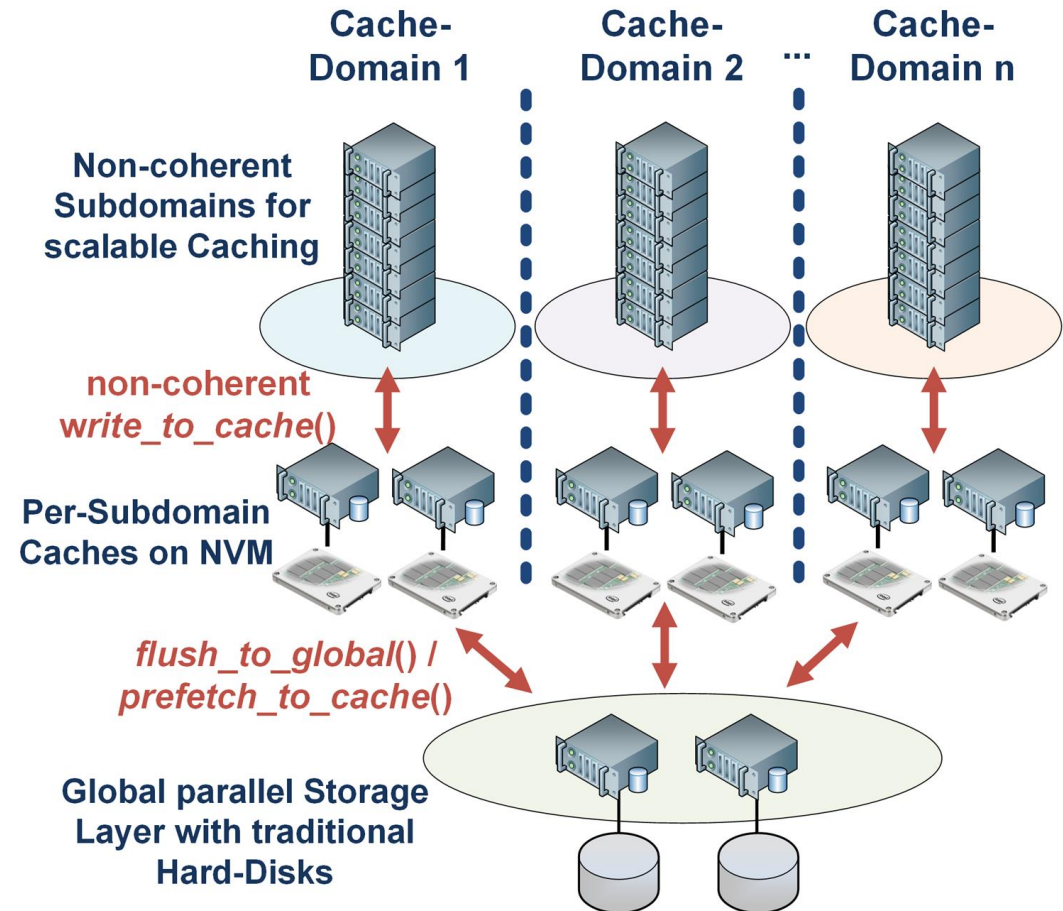
FWI (full wave inversion) code

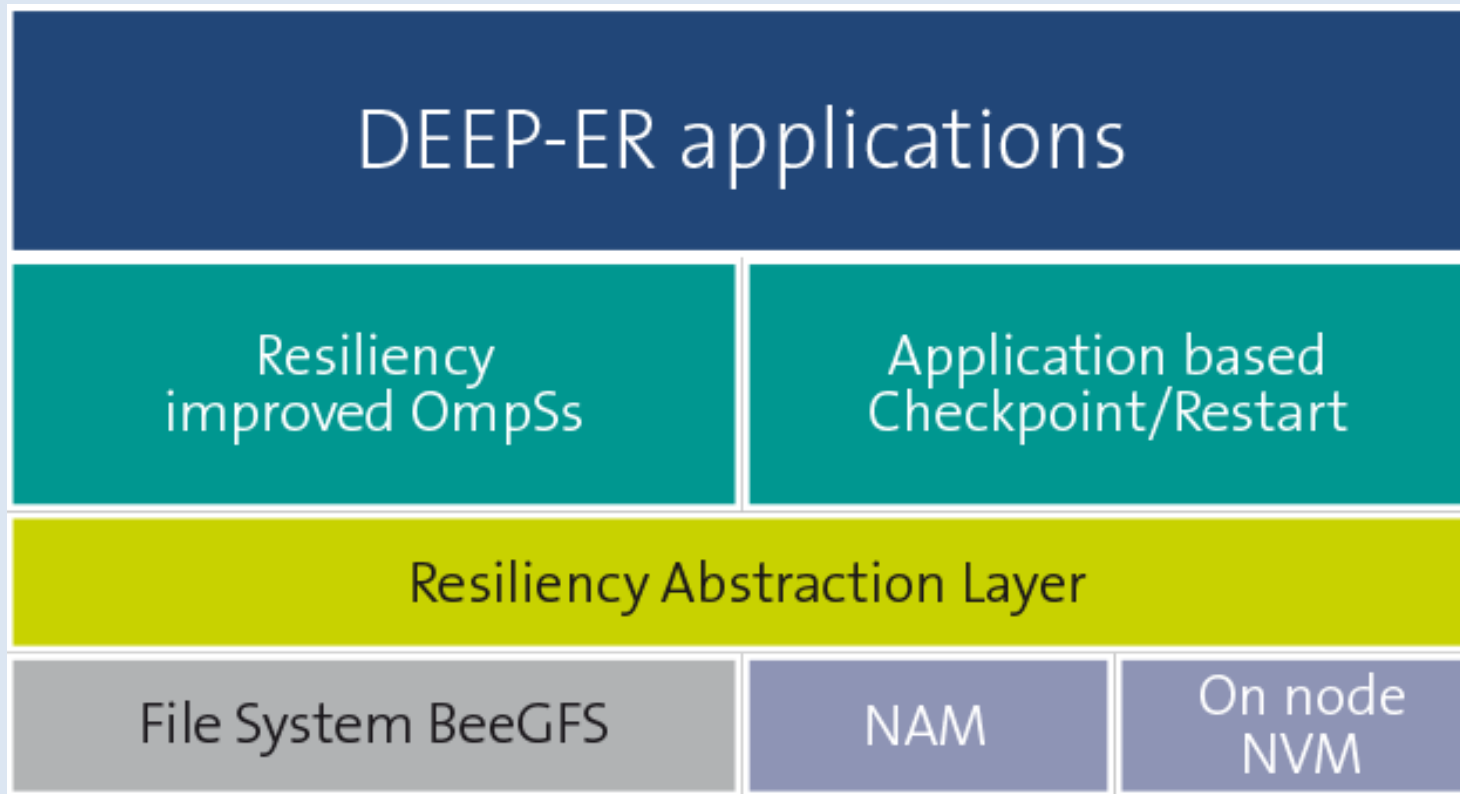




- Improve I/O scalability on all usage-levels
- Used also for checkpointing

- Two instances:
 - Global FS on HDD server
 - Cache FS on NVM at node
- API for cache domain handling
 - Synchronous version
 - Asynchronous version





- Develop a hierarchical, distributed checkpoint/restart mechanism leveraging DEEP-ER architecture

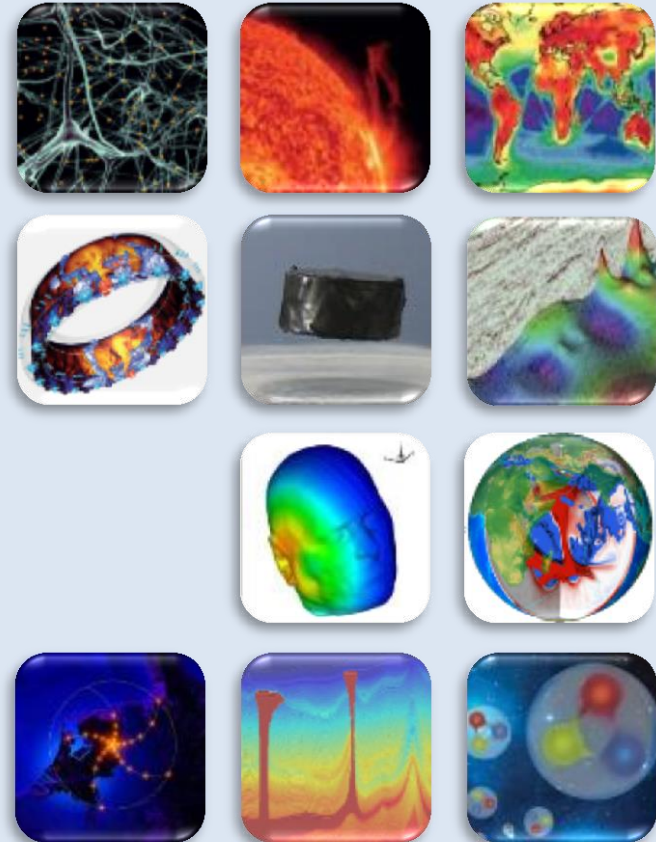
APPLICATIONS

- **DEEP+DEEP-ER applications:**

- Brain simulation (EPFL)
- Space weather simulation (KULeuven)
- Climate simulation (Cyprus Institute)
- Computational fluid engineering (CERFACS)
- High temperature superconductivity (CINECA)
- Seismic imaging (CGG)
- Human exposure to electromagnetic fields (INRIA)
- Geoscience (LRZ Munich)
- Radio astronomy (Astron)
- Oil exploration (BSC)
- Lattice QCD (University of Regensburg)

- **Goals:**

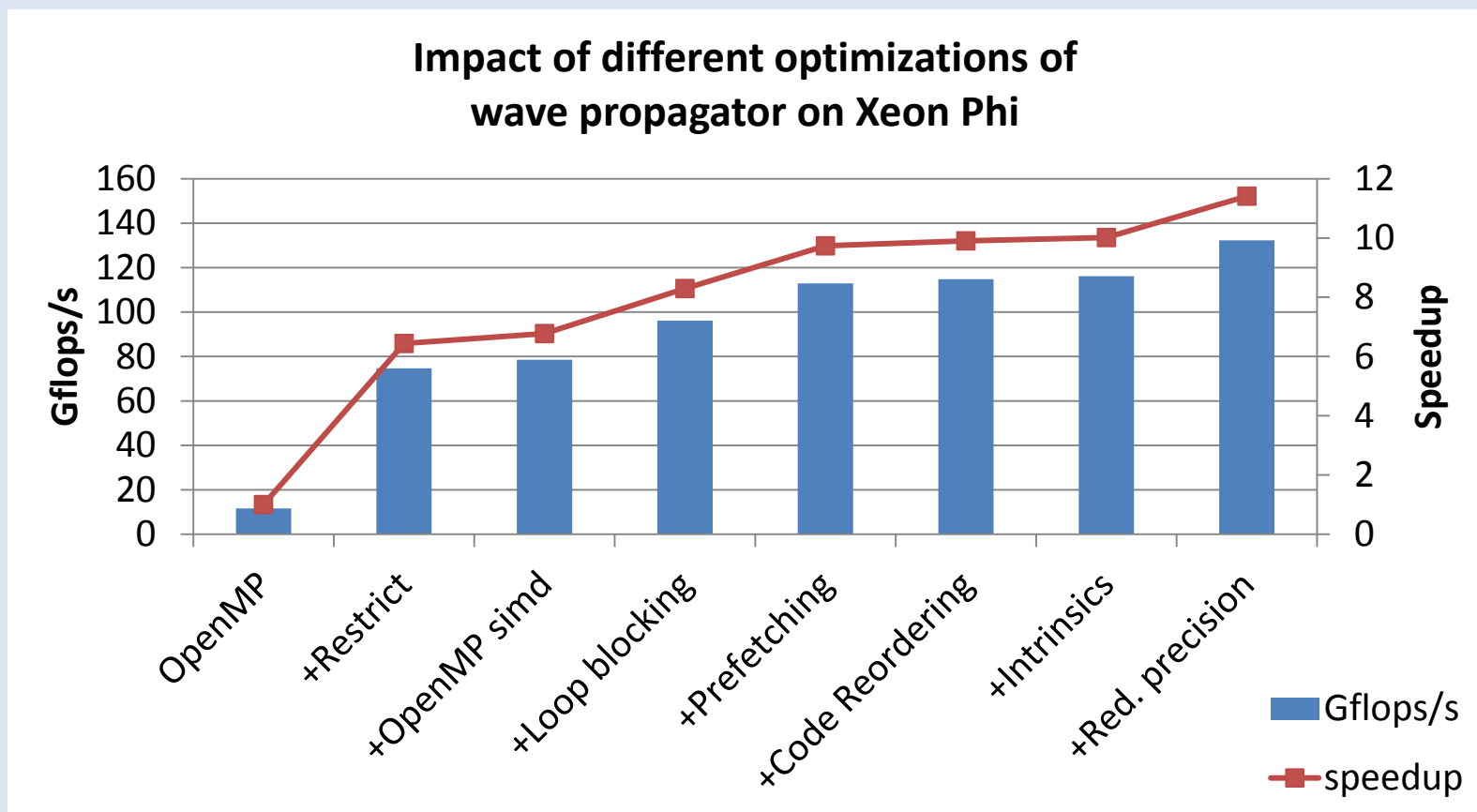
- Co-design and evaluation of architecture and its programmability
- Analysis of the I/O and resiliency requirements of HPC codes



- More **flexible** than a standard architecture
 - This enables different use models:
 1. Dynamic ratio of processors/coprocessors
 2. Use Booster as pool of accelerators (globally shared)
 3. Discrete use of the Booster
 4. Discrete use + I/O offload
 5. Specialized symmetric mode
- Enables a **more efficient use of system resources**
 - Only resources actually needed are blocked by applications
 - Dynamic allocation further increases system utilization

BSC: Enhancing Oil Exploration (FWI, wave propagator)

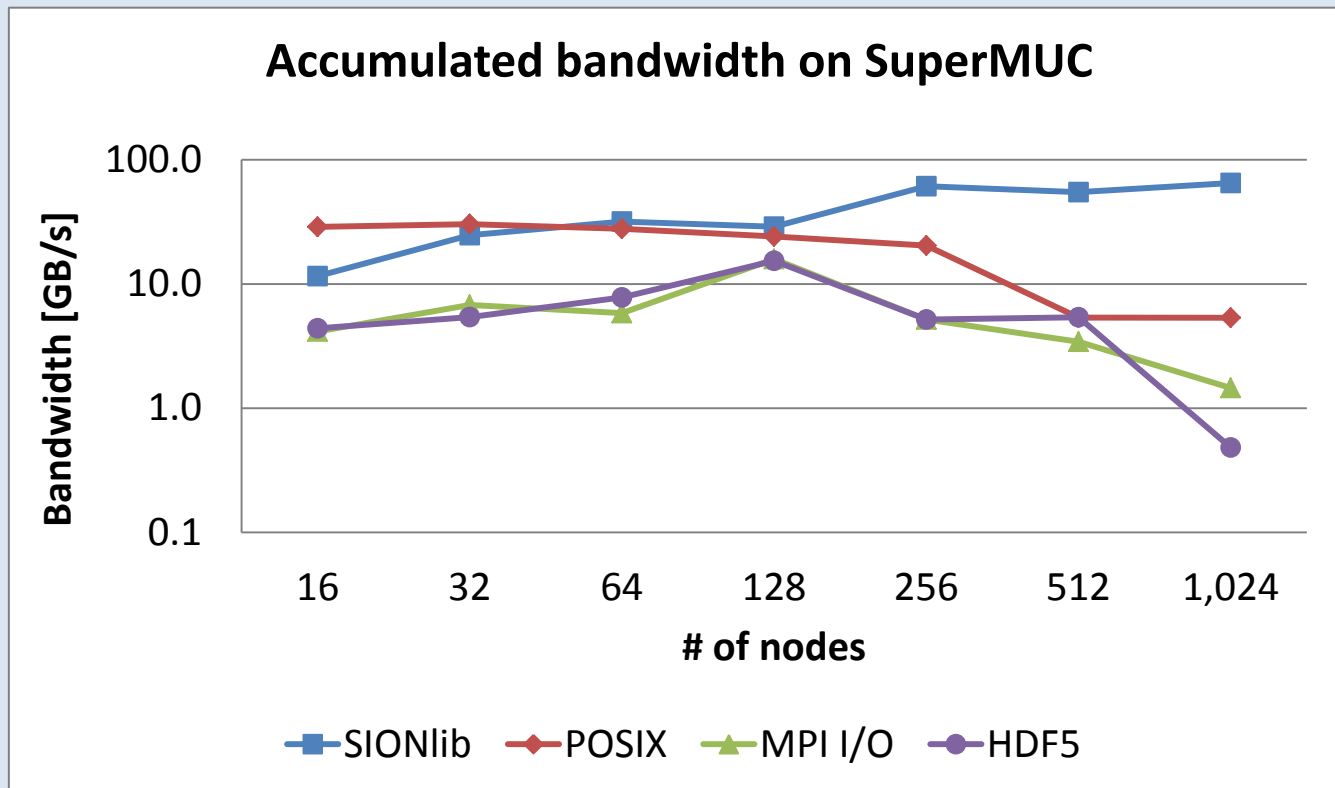
1 XeonPhi (60 cores), 180 OpenMP threads



LRZ: Rapid crustal deformation & earthquake source equation (Seisol)

1 process per node, 16 threads per process

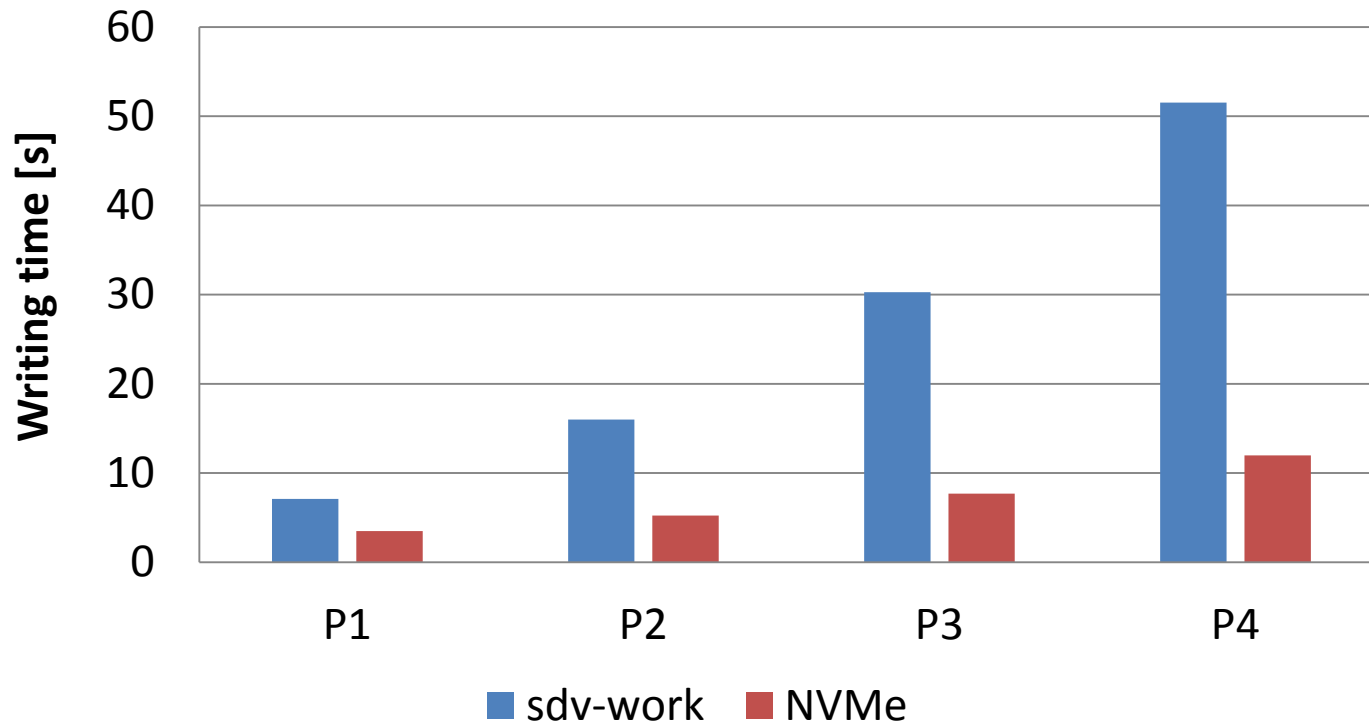
writing 20 checkpoint files (4GB/checkpoint)



Inria: Assessment of Human exposure to EM fields

24 MPI processes, 1 thread per process

I/O performance of MAXW-DGTD



Increasing
model precision
 $P1 < P2 < P3 < P4$

- **Cluster-Booster Architecture:**
 - Alternative approach to heterogeneity
 - High flexibility enabling various use modes
- **Hardware components:**
 - Booster (new kind of cluster of accelerators)
 - GreenICE Booster (2-phase immersion cooling)
 - EXTOLL network tested at scale
 - Warm-water cooling
 - Memory hierarchy based on NVM
 - Network Attached Memory

- **Software**

- **Cluster-Booster Protocol**: low-level communication protocol between different high-speed networks
- **Programming environment** for future heterogeneous systems
- **ParaStation Global MPI** supporting EXTOLL and CBP
- **OmpSs** extensions for DEEP Offload
- **Resiliency** extensions for OmpSs (task recovery) and ParaStation
- **BeeGFS** extension for local caches (on NVM)
- **SIONlib** extensions for buddy-checkpointing, integration with SCR and use of BeeGFS functionality
- **E10** scalability optimisations for MPI-I/O
- **Extrae/Paraver** support for DEEP Offload
- **Applications** modernisation and optimisation

EU-Exascale projects
20 partners
Total budget: 28,3 M€
EU-funding: 14,5 M€
Nov 2011 – Mar 2017

Visit us @
ISC'16, Frankfurt
(Germany)
20.-22.06.2016

-Booth #1340
-BoF #11
-Workshop

